DATA PAPER

A Data Collection on Secondary School Students' STEM Performance and Reading Practices in an Emerging Country

Quan-Hoang Vuong[1,2], Viet-Phuong La[1,2], Manh-Toan Ho[1,2], Thanh-Hang Pham[3], Thu-Trang Vuong[2], Ha-My Vuong[2] & Minh-Hoang Nguyen[1,2†]

[1]Centre for Interdisciplinary Social Research, Phenikaa University, Yen Nghia Ward, Ha Dong District, Hanoi 100803, Vietnam

[2]A.I. for Social Data Lab, Vuong & Associates, Dong Da District, Hanoi 100000, Vietnam

[3]School of Business and Management, RMIT Vietnam University, Hanoi 100000, Vietnam

Keywords: STEM education; Socio-economic status; Book reading; Emerging country; Students

Citation: Vuong, Q-H., et al.: A data collection on secondary school students' STEM performance and reading practices in an emerging country. Data Intelligence 3(2), 336-356 (2021). doi: 10.1162/dint_a_00091
Received: January 19, 2021; Revised: February 25, 2021; Accepted: February 25, 2021

ABSTRACT

Science, technology, engineering, and mathematics (STEM) education has become a critical factor in promoting sustainable development. Meanwhile, book reading is still an essential method for cognitive development and knowledge acquisition. In developing countries where STEM teaching and learning resources are limited, book reading is an important educational tool to promote STEM. Nevertheless, public data sets about STEM education and book reading behaviors in emerging countries are scarce. This article, therefore, aims to present a data set of 4,966 secondary school students from a school-based data collection in Vietnam. The data set comprises of five major categories: 1) students' personal information (including STEM performance), 2) family-related information, 3) book reading preferences, 4) book reading frequency/habits, and 5) classroom activities. By introducing the designing principles, the data collection method, and the variables in the data set, we aim to provide researchers, policymakers, and educators with well-validated resources and guidelines to conduct low-cost research, pedagogical programs in emerging countries.

† Corresponding author: Minh-Hoang Nguyen (Email: hoang.nguyenminh@phenikaa-uni.edu.vn; ORCID: 0000-0002-7520-3844).

1. INTRODUCTION

Globally, technology has become a crucial component of our cultural, social, and economic landscape. Scientific and technological innovation is essential means to promote economic advancement and social justice. However, the emerging countries, which are lagged economically, are also markedly inferior in their research and technology developments [1,2,3]. Therefore, it is suggested that a focus on science, technology, engineering, and mathematics (STEM) education may serve as a useful direction for these nations. It is also a crucial factor to achieve United Nations Educational, Scientific and Cultural Organization (UNESCO)'s Sustainable Development Goal (SDG) 4—Quality Education [4]. In particular, STEM education can equip young people with the skills to adapt to the digital age and create life-changing opportunities for themselves and their families.

To improve STEM learning effectiveness, especially in developing contexts, reading is a critical element that needs addressing in the era of technology. Reading is a fundamental activity that helps enhance various cognitive capabilities, academic achievement, and educational and occupational attainment in young adulthood [5,6,7]. Children's technological problem-solving skills could be improved by developing a reading culture at home [7]. Furthermore, reading is also found to facilitate the creation of new ideas and inquiries for the problem-solving process involved in STEM learning [8]. Empirically, Braun et al. [9] and Fang and Wei [10] found that secondary school students in the United States who spend time reading additional books perform better in both reading and scientific subjects than those who do not. However, a majority of prior studies have been conducted in developed settings. In contrast, the topics regarding STEM performance and reading behaviors among children in developing countries remain understudied.

The lack of data might induce the research shortage in emerging countries. Observing multiple open globally-scoped data repositories (Mendeley Data, Zenodo, Harvard Dataverse, Figshare, etc.) and data journals (e.g., *Scientific Data* and *Data in Brief*), we found limited data sets on STEM education and reading behaviors among children. There are series of national survey data concerning children's educational conditions and performance. Still, the majority of them were conducted in developed countries, specifically Australia and United States [11,12].

Of those in developing contexts, a data set provided by Ranjeeth et al. [13] offers resources to investigate the relationship between Indian secondary school students' academic performance and outside-class habits. Nonetheless, it lacks focus on STEM-related achievement, and its questions are not systematically structured. Another set of data about Nigerian students' academic performance in STEM-related disciplines is also accessible on *Data in Brief* [14]. The contents are collected solely from undergraduate schools. Besides data sets made open on data repositories and journals, we have found a wide array of global data sets directly related to the SDG 4 on UNESCO Institute for Statistics's online database [15]. Those data sets are structured to scope the macro progress in achieving SDG 4's targets, whereas data sets concentrating on individual-level development are scarce.

Therefore, the current paper presents a highly organized and multifaceted data set regarding socio-economic status, family conditions, reading habits and preferences, perceived classroom activities, and academic performance among junior high school students in Vietnam. Vietnam is a unique case of an emerging nation that has been highly regarded for K–12 students' STEM achievement. Notably, Vietnam's students have recorded high-performance levels in the Programme for International Student Assessment (PISA) over the three years in which they participated [16,17,18], despite recent participation. When putting this in alignment with Vietnamese households' comparatively limited socio-economic status, this significant attainment is a worth investigating case.

The current data set is a valuable resource to provide evidence that helps policymakers and education leaders to improve educational quality and equity. Given the low cost of a substantial data collection [19] and the shortage of earlier research on similar issues in emerging countries, the current data set and its design might be useful points of reference for other developing countries with similar context with Vietnam.

Several high-quality findings regarding book reading and STEM performance have been obtained and published using only a portion of the current data set [20,21,22]. A typical example is a recent publication that analyzes the social gap and gender disparity in Vietnam, which associates with STEM outcomes [20]. As the reproducibility crisis occurs across disciplines and hampering knowledge accumulation progress, Munafò et al. [23] encourage scientists to improve transparency through open science. Thus, the open access to the current data set will enable other researchers to replicate and validate the published findings, strengthening the results' integrity and reliability amidst the movement to make science more transparent [24,25].

2. METHODOLOGY

2.1 Data Collection Sample and Design

The current research's data [26] were collected through a school-based data collection of STEM performance, reading habits, and preferences of junior high school students (Grade 6 through Grade 9, which corresponds to age 11 to 15) in Ninh Binh Province, Vietnam. The data collection was conducted in two periods. The former was between December 2017 and January 2018, while the latter was from February until July 2018. In total, 4,966 responses of adolescent students were acquired.

The questions in the questionnaire were initially designed by Vuong and Associates office. Then, an officer in the provincial Department of Education and Training distributed the paper questionnaires to academic administrators of 16 public junior high schools in the area who were later in charge of gathering the questionnaire responses. All participants in the data collection process adhered to the institutions' ethical code responsible for the research.

The research project can be structured into five phases: 1) questionnaire design, 2) questionnaire collection, 3) quality control for questionnaire answers, 4) data set generation, and 5) data analysis.

2.2 Data Validation

For ensuring the quality and validity of the data, the data were rigorously checked several times in three phases of the data collection: 1) in-school phase, 2) gathering phase, and 3) computerizing phase (Figure 1). During the first phase, the homeroom teachers were in charge of collecting the questionnaires from students and initially checking the response accuracy and missing answers; then, the person who took care of the school academic activities was responsible for validating the collected questionnaires one more time. In the next phase, all the questionnaires from 16 schools were transferred to the data collection administrator—an officer in the provincial Department of Education and Training, for the third quality and validity check. After the examination, the administrator handed the responses over to the current study's principal investigator for the computerizing phase. During this phase, research team members entered the data into an MS Excel file and cross-checked the file to ensure that the inputted data precisely represented the answers on questionnaires. When there was doubt about the response, the team members consulted with the principal investigator for resolution. The research team contacted the data collection administrator for validation in case there existed serious errors.

The data set has been capitalized using both frequentist and Bayesian approaches in several publications, including [20,21,22]. From these works, various significant results have been found. Regarding socio-economic backgrounds, gender had a negligible correlation with students' STEM results at schools, and it was also found that female students can achieve better results than their male counterparts [18]. Regarding the correlation with reading behavior, empirical results also show that higher grades in STEM-related subjects are predicted by reading interest, with students who love reading books achieving higher scores than those who take no interest in books [19]. For raising the reading interest among students, scholarly culture at home and reading practices were suggested to be significant determinants [20].

In the current data descriptor, we also conduct an exemplary analysis using the current data set utilizing the Bayesian estimation and Markov Chain Monte Carlo (MCMC) technique. The example is shown in Section 3 below.

2.3 Data Collection

Questionnaires were distributed to all classes in 16 public junior high schools by the academic administrators and then collected by the homeroom teachers. Eventually, 4,966 students' responses were received. Overall, boys' and girls' responses shared half of the sample (49.36% and 49.62%, respectively), while the remaining 1.03% (51/4,966) of the total students did not report their biological sex. The percentage of students among four grades (from Grade 6 to Grade 9) was proportionately similar with 24.90% (1,237/4,966) from the sixth grade, 24.04% (1,194/4,966) from the seventh grade, 23.86% (1,185/4,966) from the eighth grade, and 25.51% (1,267/4,966) from the ninth grade. A majority of the sample reported to like reading (88.40%), and 41.80% of students were the first child in their families. Students' average score of STEM-related subjects ("APS45") and favorite type of books ("Topic") are plotted in Figure 2.

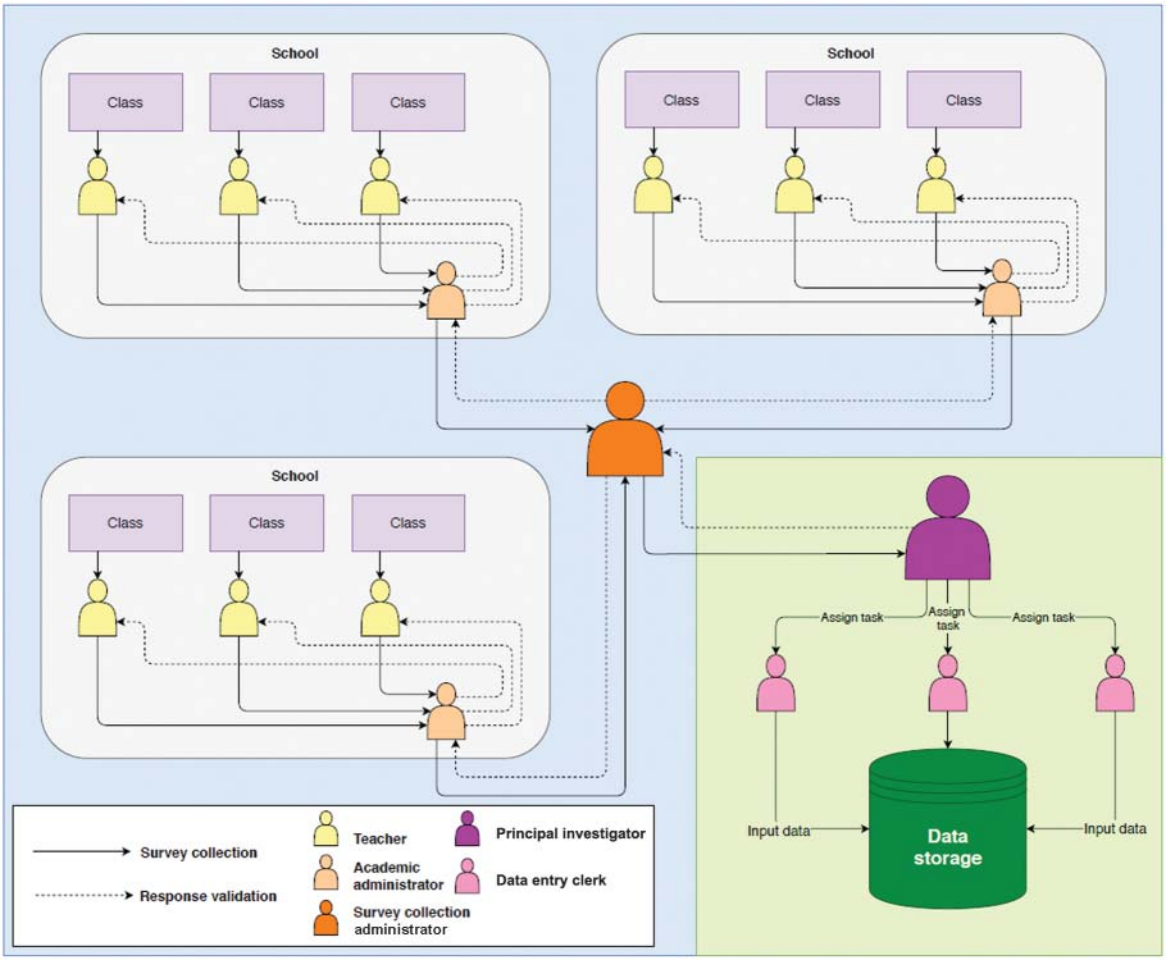

Figure 1. Diagram of the data collection and validation. Note: There were three primary data collection phases: 1) in-school phase, 2) gathering phase, and 3) computerizing phase. The questionnaires' responses were checked one or two times before being proceeded to the next phase. In case there was doubt regarding the nature of an answer, the data validator would check with the former validator in the process.

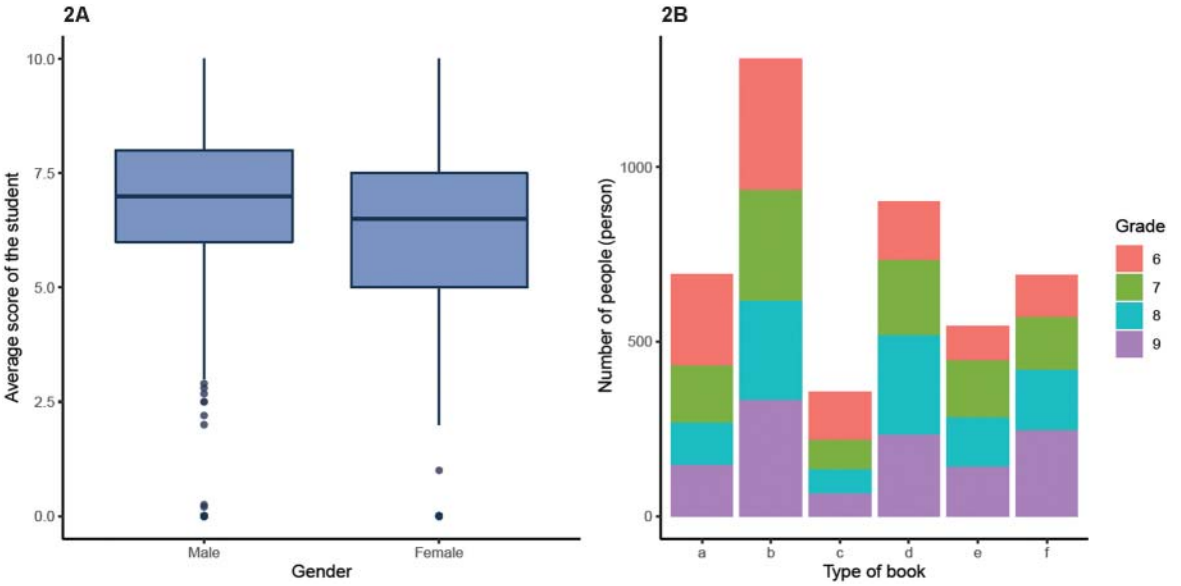

Figure 2. Visual representations of students' information. Note: Figure 2A displays the distribution of the average score of the most recent 45-minute tests of Mathematics, Physics, Chemistry, and Biology by the sex of the students. Overall, male students had a higher mean score of STEM-related subjects than female students. Figure 2B shows the distribution of students by their favorite type of book, and within each favorite type of book the distribution of students by their grade. "a", "b", "c", "d", "e", and "f" correspond with mathematics/physics, literature, foreign language, natural science/chemistry/biology, history/geography, and information technology, respectively. 27.57% (1,369/4,966) of students were keen on literature-related books, and the distribution of students by grade within the literature type of book was generally balanced.

2.4 Response Coding

After receiving the hand-written questionnaires, the research team codified responses into 37 different variables, which could be classified into five main categories: 1) personal information (including STEM performance), 2) family-related information, 3) book reading preferences, 4) book reading frequency/habits, and 5) classroom activities. The variables' coding and a brief description of Categories 1 and 2 are shown in Table A1, while those of Categories 3 to 5 are displayed in Table A2. Below is a detailed explanation of each category's variables in the text.

1) Personal information. The personal information category includes "Sex" (male, female), "Grade", "School", "APS45", "APSVNEN", "FutureJob", and "Hobby". The variable "APS45" represents the average scores of the most recent 45-minute examinations of Mathematics, Physics, Chemistry, and Biology. In contrast, the variable "APSVNEN" indicates the average score of mid-term tests of Mathematics and other natural science subjects. As for "Hobby", students were given six main kinds of activities, namely: reading books, watching TV/listening to music, housework/farming, observing nature, interacting with friends/family members, and others. For better utility, these activities were coded as "a", "b", "c", "d", "e", and "f", respectively.

2) Family-related information. The second category regarding family-related information could be subsequently classified into three sub-groups. In the first sub-group, "RankingF" and "NumberofChi" demonstrate the student's birth order in the family and the number of children in the student's family, respectively. The second sub-group regards parents' characteristics, such as father's/mother's education ("EduFat"/"EduMot"), father's/mother's age ("AgeFat"/"AgeMot"), and father's/mother's career ("CareerFat"/"CareerMot"). The educational level of the student's parents was captured in four levels: under high school ("UnderHi"), high school ("Hi"), undergraduate ("Uni"), and graduate school ("PostGrad"). The last sub-group focuses on the economic aspects of the family, including perceived economic status ("EcoStt"), awareness of the family's monthly income ("KnowledgeInc"), and estimated monthly income ("EstIncome"). The perceived economic status was coded as follows: "poor" = low-income family, "med" = medium-income family, and "rich" = wealthy family, whereas the variable "KnowledgeInc" only consists of two answers, "yes" or "no". The students were asked to report their families' estimated income ("EstIncome") in Vietnam Dong.

3) Book reading preferences. The third category primarily concerns the reading preferences of students, which could be shown by seven variables. The reading interest of a student was referred to as "Readbook" (yes, no). Favourite type of book was captured as the "Topic" (mathematics/physics = "a", literature = "b", foreign language = "c", natural science/chemistry/biology = "d", history/geography = "e", and information technology = "f"). Besides the favorite type of book, the prioritized book type when being gifted was represented by a categorical variable "Typebook" (novel = "a", biography = "b", popular science = "c", arts = "d", vocational instruction = "e", and other = "f"). Subsequently, the reason of the prioritization was shown by the "Reason" (personal preferences = "a", recommended by parents = "b", recommended by teachers/friends = "c", and serendipity = "d"). The "PrioAct" specifies the primary activity conducted by students when meeting a good book; it consists of "a" (sharing with friends/family), "b" (recording), "c" (applying the content to daily life), and "d" (reflecting and relating to personal knowledge). Students' prime activity after reading a good book was shown as the "AftAct" (finding more books on the exact issue = "a", finding more books on the related issue = "b", finding books on the new issue = "c", and reading the book again = "d"). Eventually, the student was asked to provide two favorite books, of which the variable was recorded as the "Read_like".

4) Book reading frequency/habits. The fourth category comprises five variables respecting the reading habits of students. The first two variables, "TimeSci" and "TimeSoc", concern the length of time students spent reading science books and literature/social sciences-related books daily, respectively. Both variables were coded as follows: less than 30 minutes = 1, between 30 and 60 minutes = 2, and over one hour = 3. Two other variables in this category are habits conducted by students' parents, such as buying books ("Buybook") and reading stories ("Readstory") for their children. These two variables were constructed with binary responses – "yes" or "no". The last variable in this category was generated from the question regarding books' primary source ("Source"). To answer this question, students were asked to choose one alternative among buying books on their own or with parents' money ("buy"), borrowing from friends or libraries ("borrow"), or being gifted or rewarded ("gift").

5) Classroom activities. The classroom is an important place to build up students' reading behaviors, interests, and preferences. Therefore, we created a separate section with four questions in the questionnaire to examine students' perceptions of classroom activities. The first question, which corresponds to the binary variable "EncourAct" (yes or no), is to see whether students were keen on activities that encouraged reading. Then, students were asked to give the activity that they were most interested in ("MostlikedAct") among the four following options: book exhibition = "a", storytelling competition = "b", story-writing competition = "c", and illustrating books' content by drawing = "d". Besides favorite activities that encouraged reading, we also examined students' perceptions of bookshelf conditions in the classroom ("Bookcase"). Students were required to assess the condition according to the following alternatives: diverse and interesting = "a", lack of good titles = "b", lack of book = "c", and no bookshelf = "d". The last variable ("Notread_like") was generated from an open-ended question in which students could freely provide three favorite titles that they would like to read but were not available on the classroom bookshelf.

3. EXEMPLARY DATA ANALYSIS

The responses were initially inputted into an MS Excel file, and then converted to a comma-separated values (.csv) file. Bayesian linear regression applying the Markov Chain Monte Carlo (MCMC) technique was employed for analyzing the data with the dependent variable being "APS45" (average score of the most recent 45-min tests of Math, Physics, Chemistry and Biology of students) and independent variables being "Readbook" (reading interest) and "Topicgr" (students' perceived economic status). Here, the Bayesian approach is preferable rather than the frequentist approach because of several replication issues induced by the *P*-value. Specifically, "the wide sample-to-sample variability" in the *P*-value and the *P*-hacking practices are two major contributors among reasons that exacerbate the reproducibility crisis [23,27,28].

The procedure was adapted from the Bayesian protocol for examining social data by Vuong et al. [29]. All the priors of the model's coefficients were set as uninformative before simulation. The bayesvl R package was utilized due to its user-friendly operation and graphical visualization power [30,31].

The Bayesian analytical model to examine the associations between "APS45" and "Readbook" and "Readstory" is defined as follows:

$$O_{APS45} \sim \alpha + \text{Readbook} + \text{Readstory}$$

The codes utilized in the R (version 4.0.2) for the Bayesian modelling and MCMC simulation are shown below:

```
# Data preparation
data<-read.csv("C:/…/STEM5000.csv",header = T,stringsAsFactors = TRUE)
keeps <- c("APS45", "Readbook", "Readstory")
data <- data[keeps]
data<- na.omit(data)

require(dplyr)
dat$Readbook <-
  as.numeric(
    case_when(
      dat$Readbook %in% c("yes") ~ 1,
      dat$Readbook %in% c("no") ~ 0
    )
  )

dat$Readstory <-
  as.numeric(
    case_when(
      dat$Readstory %in% c("yes") ~ 1,
      dat$Readstory %in% c("no") ~ 0
    )
  )

# Model construction
library(bayesvl)

model<-bayesvl()
model<-bayesvl()
model<-bvl_addNode(model,"APS45","norm")
model<-bvl_addNode(model,"Readbook","binom")
model<-bvl_addNode(model,"Readstory","binom")


model <- bvl_addArc(model, "Readbook", "APS45", "slope")
model <- bvl_addArc(model, "Readstory", "APS45", "slope")

# Stan code generation
model_string <- bvl_model2Stan(model)
cat(model_string)

# Model fit
model <- bvl_modelFit(model, data, warmup = 2000, iter = 5000, chains = 4, cores = 4)
summary(model)

# Visualization of posterior estimates' trace plots
bvl_plotTrace(model)

# Visualization of posterior estimates' Gelman plots
bvl_plotGelmans(model, NULL, 1, 3)

# Visualization of posterior estimates' density distribution
bvl_plotDensity2d(model,"b_Readbook_APS45","b_Readstory_APS45",color_scheme = "orange") + theme_
bw()
```

The posterior coefficients of the estimation of "APS45" against "Readbook" and "Readstory" are displayed in Table 1. Table 1 presents the posterior estimates acquired from the Bayesian linear regression model employing "APS45" as a dependent variable and "Readbook" and "Readstory" as independent variables. Two independent variables were set as binary variables during the model fitting process with "yes" = 1 and "no" = 0. The model was fitted with four Markov chains, 5,000 iterations, and 2,000 warm-up iterations. After running the MCMC simulation, we obtain a good demonstration of convergence by two standard diagnostics: All coefficients' Rhat's values (Gelman shrink factor) were one, and all values of n_eff (effective sample size) were beyond 1,000 (Table 1). The visual diagnostic of Markov chains' convergence is also displayed in Figures 3 and 4. "Readbook" and "Readstory" coefficients obtained positive mean and acceptable standard deviation ($\mu_{\text{Readbook}} = 0.44$ and $\sigma_{\text{Readbook}} = 0.14$; $\mu_{\text{Readstory}} = 0.14$ and $\sigma_{\text{Readstory}} = 0.09$). Based on these results, we suggest that improving the interest in reading books and encouraging parents to read stories for their children might result in better STEM performance among students. The coefficients' posterior distributions are displayed in Figures 5A and 5B, the histogram and two-dimensional density plot, respectively.

Table 1. Estimated posterior coefficients.

Coefficient	Mean (μ)	Standard deviation (σ)	n_eff	Rhat
Constant	5.66	0.14	4572	1
"Readbook"	0.44	0.14	4570	1
"Readstory"	0.16	0.09	7996	1

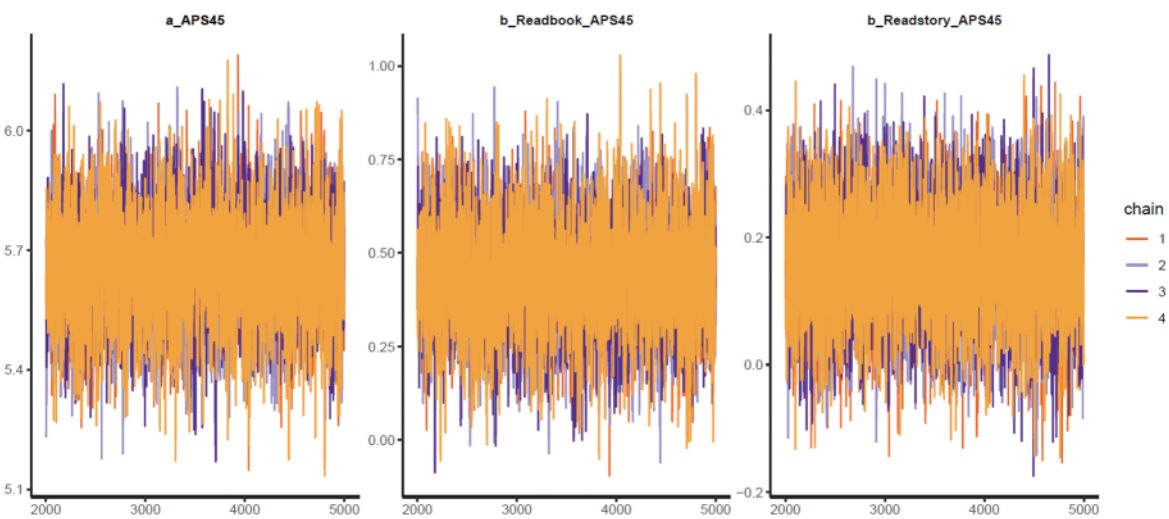

Figure 3. Trace plots of MCMC draws of coefficients. Note: Figure 3 illustrates the Markov property's visual diagnostic of the coefficients' posterior estimates by trace plots. The trace plots in Figure 3 were created from four Markov chains, each containing 5,000 iterations, of which 2,000 were warm-up iterations. All the trace plots exhibit two primary characteristics meeting the Markov property: stationarity and good mixing.

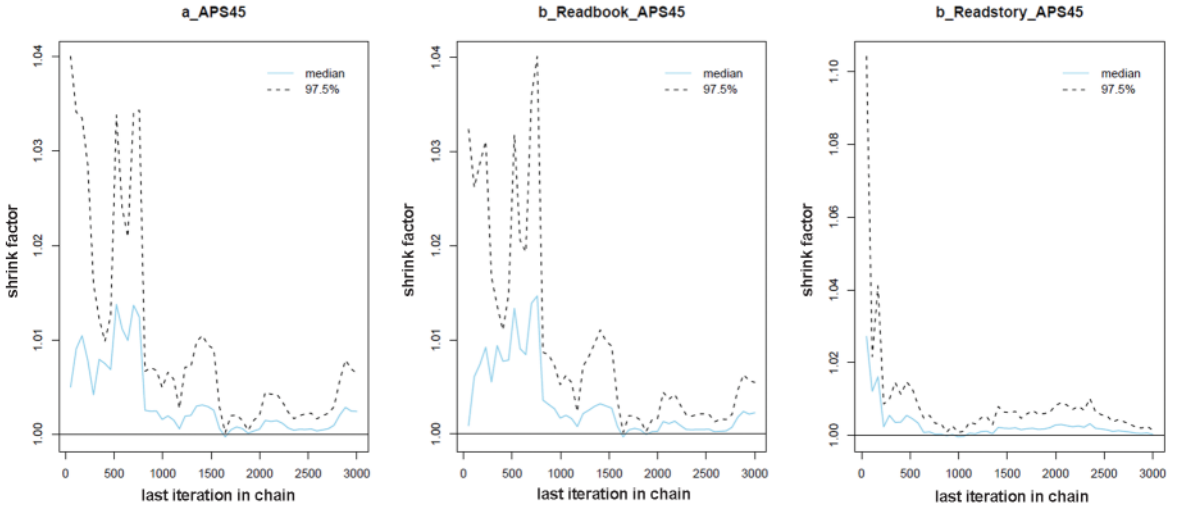

Figure 4. Gelman plots of coefficients. Note: Figure 4 displays the change of the scale-reduction factor of each coefficient over time. The factors go down and up at first but gradually drop to below 1.01 after the warm-up period (or the 2,000th iteration), showing the good convergence of coefficients' Markov chains.

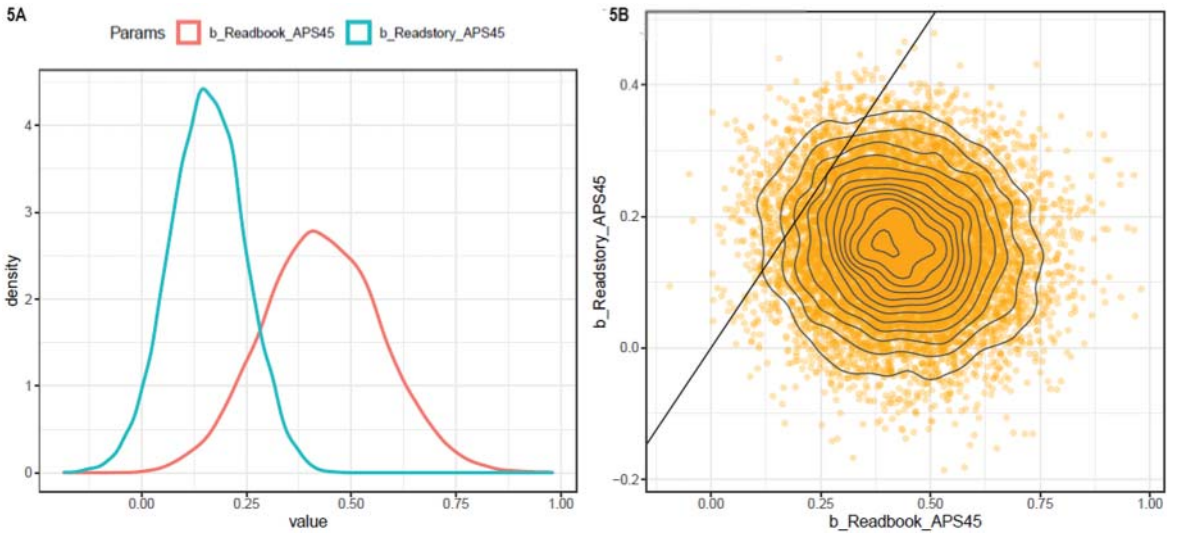

Figure 5. Two-dimensional density plot of posterior estimates. Note: Figure 5A illustrates the posterior distribution of coefficients' "Readbook" and "Readstory" on a line histogram, while Figure 5B displays the distribution of simulated posterior values on a two-dimensional surface. In both Figures 5A and 5B, coefficients' distributions mostly correspond with the axes' positive values, showing the positive influence of "Readbook" and "Readstory" on students' STEM performance ("APS45"). The coefficient "Readbook" had a more significant impact on "APS45" than "Readstory".

4. USAGE NOTES AND CONCLUSION

The available data set offers several key features that can be used in future research. First, researchers can continue studying the relationship between children's reading habits and educational achievements. The effects of parental involvements, gender, and school activities on children's reading behaviors in an emerging context are also worth exploring. Evidence from these analyses will be beneficial to educators and policymakers, especially those in developing countries.

The deposition of the data set also allows researchers to replicate previous findings. Amidst the reproducibility crisis [23,27,28], replication is an important aspect to produce robust and precise evidence. Making the data set open also improves the transparency and published results' integrity and reliability. Thus, previous findings will be put under the scrutiny of post-publication review. Hopefully, this practice will support the cause of open access and prevent irreproducibility.

Finally, while the data set is an important asset, the survey's design principles are also valuable. We designed the study on a low-cost basis due to the limited fundings. The data set and its subsequent analysis have somewhat proved that this method is cost-effective. Thus, researchers from developing countries and those working with limited resources can use these design principles. In a larger context, we hope this method will also reduce the cost of science [19] and contribute significantly to scientific communities worldwide.

In conclusion, the current data set was systematically designed and collected. Hence, it provides researchers, policymakers, and educators with well-validated resources regarding secondary students' multifaceted aspects for formulating pedagogical programs and strategies in Vietnam and other countries, especially emerging countries with similar contexts.

AUTHOR CONTRIBUTIONS

Q.-H. Vuong (hoang.vuongquan@phenikaa-uni.edu.vn) designed the questionnaire, administered the data collection, and validated the data set. V.-P. La (phuong.laviet@phenikaa-uni.edu.vn) validated and supervised the data input process. T.-T. Vuong (thutrang.vuong@sciencespo.fr) and H.-M. Vuong (vuonghamy2003@gmail.com) contributed to the data input and validation process. T.-H. Pham (hang.pt.88@gmail.com) contributed to manuscript writing. M.-T. Ho (toan.homanh@phenikaa-uni.edu.vn) contributed to data input, validation process, and manuscript writing. M.-H. Nguyen (hoang.nguyenminh@phenikaa-uni.edu.vn) contributed to manuscript writing and software implementation.

ACKNOWLEDGEMENTS

We would like to send our gratitude to the research staff of Vuong & Associates (Hanoi, Vietnam) for assisting in collecting data, especially Do Thu Hang, Ho Manh Tung, Nguyen To Hong Kong, and Dam Thu Ha. Our most sincere appreciations also go on to personnel of junior high schools and provincial departments that provided support during the data collection.

DATA AVAILABILITY STATEMENT

All the data were anonymized and stored in a .csv format file that are available in the Science Data Bank repository, https://doi.org/10.11922/sciencedb.j00104.00090, under an Attribution 4.0 International (CC BY 4.0).

CODE AVAILABILITY STATEMENT

Data were analyzed and visualized using the open statistical software R (version 4.0.2). All the code employed for generating figures and conducting Bayesian analysis was included in a PDF file available at https://doi.org/10.11922/sciencedb.j00104.00090.

REFERENCES

[1] Arauco, V.P., et al.: Explaining low redistributive impact in Bolivia. Public Finance Review 42(3), 326–345 (2014)
[2] Confraria, H., Godinho, M.M.: The impact of African science: A bibliometric analysis. Scientometrics 102, 1241–1268 (2015)
[3] Hien, P.D.: A comparative study of research capabilities of East Asian countries and implications for Vietnam. Higher Education 60, 615–625 (2010)
[4] United Nations. Sustainable development goals—Quality Education. Available at: https://www.un.org/sustainabledevelopment/education/ (2020). Accessed 25 February 2021
[5] Ritchie, S.J., Bates, T.C., Plomin, R.: Does learning to read improve intelligence? A longitudinal multivariate analysis in identical twins from age 7 to 16. Child Development 86(1), 23–36 (2015)
[6] Rabiner, D.L., Godwin, J., Dodge, K.A.: Predicting academic achievement and attainment: The contribution of early academic skills, attention difficulties, and social competence. School Psychology Review 45, 250–267 (2016)
[7] Evans, M., et al.: Scholarly culture and occupational success in 31 societies. Comparative Sociology 14(2), 176–218 (2015)
[8] Pearson, P.D., Moje, E., Greenleaf, C.: Literacy and science: Each in the service of the other. Science 328(5977), 459–463 (2010)
[9] Braun, H., et al.: Exploring what works in science instruction: A look at the eighth-grade science classroom. (Educational Testing Service, 2009). Available at: https://files.eric.ed.gov/fulltext/ED507837.pdf. Accessed 25 February 2021
[10] Fang, Z., Wei, Y.: Improving middle school students' science literacy through reading infusion. The Journal of Educational Research 103(4), 262–273 (2010)
[11] Australian Government Department of Education, Skills and Employment. Longitudinal surveys of Australian youth, 2015 cohort (Version 4.0). Dataverse. Available at: http://dx.doi.org/10.4225/87/PJO7GB (2017). Accessed 25 February 2021
[12] United States Department of Education. National Household Education Survey, 2005 (ICPSR 4599). National Center for Education Statistics. Available at: https://doi.org/10.3886/ICPSR04599.v1 (2007). Accessed 25 February 2021
[13] Ranjeeth, S., Latchoumi, T.P., Paul, P.V.: Role of gender on academic performance based on different parameters: Data from secondary school education. Data in Brief 29, 105257 (2020)

[14] John, T.M., et al.: The role of gender on academic performance in STEM-related disciplines: Data from a tertiary institution. Data in Brief 18, 360–374 (2018)
[15] UNESCO Institute for Statistics. Sustainable Development Goal 4. (n.d.). Available at: http://data.uis.unesco.org/. Accessed 25 February 2021
[16] OECD. PISA 2015 - Results in Focus. (2018). Available at: https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf. Accessed 25 February 2021
[17] OECD. PISA 2012 - Results in Focus (2014). Available at: http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview.pdf. Accessed 25 February 2021
[18] VNS. VN gets high scores but not named in PISA 2018 ranking. Available at: https://vietnamnews.vn/society/569454/vn-gets-high-scores-but-not-named-in-pisa-2018-ranking.html (2019). Accessed 25 February 2021
[19] Vuong, Q.-H. The (ir)rational consideration of the cost of science in transition economies. Nature Human Behaviour 2, 5 (2018)
[20] Ho, M.-T., et al.: An analytical view on STEM education and outcomes: Examples of the social gap and gender disparity in Vietnam. Children and Youth Services Review 119, 105650 (2020)
[21] Le, T.-T.-H., et al.: Reading habits, socio-economic conditions, occupational aspiration and academic achievement in Vietnamese junior high school students. Sustainability 11, 5113 (2019)
[22] Tran, T., et al.: The relationship between birth order, sex, home scholarly culture and youths' reading practices in promoting lifelong learning for sustainable development in Vietnam. Sustainability 11, 4389 (2019)
[23] Munafò, M.R., et al.: A manifesto for reproducible science. Nature Human Behaviour 1, 1–9 (2017)
[24] Vuong, Q.-H.: Reform retractions to make them more transparent. Nature 582, 149 (2020)
[25] Vuong, Q.-H.: Open data, open review and open dialogue in making social sciences plausible. Scientific data updates. Available at: http://blogs.nature.com/scientificdata/2017/12/12/authors-corner-open-data-open-review-and-open-dialogue-in-making-social-sciences-plausible/ (2017). Accessed 25 February 2021
[26] Vuong, Q.-H., et al.: A dataset on STEM performance and reading practices among secondary school students in Vietnam. Available at: https://doi.org/10.11922/sciencedb.j00104.00090 (2021). Accessed 25 February 2021
[27] Halsey, L.G., et al.: The fickle P value generates irreproducible results. Nature Methods 12, 179–185 (2015).
[28] John, L.K., Loewenstein, G., Prelec, D.: Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23(5), 524–532 (2012)
[29] Vuong, Q.-H., et al.: Bayesian analysis for social data: A step-by-step protocol and interpretation. MethodsX 7, 100924 (2020)
[30] Vuong, Q.-H., et al.: Improving Bayesian statistics understanding in the age of Big Data with the bayesvl R package. Software Impacts 4, 100016 (2020).
[31] La, V.-P., Vuong, Q.-H.: bayesvl: Visually learning the graphical structure of Bayesian networks and performing MCMC with "Stan". The Comprehensive R Archive Network (CRAN). Available at: https://cran.r-project.org/web/packages/bayesvl/index.html, version 0.8.5 (2019). Accessed 25 February 2021

APPENDIX A

Table A1. Description of variables related to personal and family information.

Category	Variable	Variable description	Variable level(s)	Code/Data type	Number (%)
1. Personal information	*Sex*	Biological sex	Male	1	2,451 (49.36%)
			Female	2	2,464 (49.62%)
	Grade	Grade	Sixth grade	6	1,237 (24.91%)
			Seventh grade	7	1,194 (24.04%)
			Eighth grade	8	1,185 (23.86%)
			Ninth grade	9	1,267 (25.51%)
	School	School's name	Open answer	Discrete data	N/A
	APS45	The average scores of the most recent 45-minute examinations of Mathematics, Physics, Chemistry, and Biology	Open answer	Interval data	From 0 to 10
	APSVNEN	The average score of midterm tests of Mathematics and other natural science subjects	Open answer	Interval data	From 0 to 10
	FutureJob	The student's favorite future job	Open answer	Discrete data	N/A
	Hobby	The student's hobby	Reading books	a	1,123 (22.61%)
			Watching TV/listening to music	b	2,093 (42.15%)
			Housework/farming	c	540 (10.87%)
			Observing nature	d	145 (2.92%)
			Interacting with friends/ family members	e	332 (6.69%)
			Others	f	682 (13.73%)

Category	Variable	Variable description	Variable level(s)	Code/Data type	Number (%)
2. Family-related information	*RankingF*	The student's birth order in the family	Open answer	Interval data	N/A
	NumberofChi	The number of children in the student's family	Open answer	Interval data	N/A
	EduFat	The student's father's educational level	Under high school	UnderHi	2,846 (57.31%)
			High school	Hi	1,308 (26.34%)
			Undergraduate	Uni	271 (5.46%)
			Graduate school	PostGrad	93 (1.87%)
	AgeFat	The student's father's age	Open answer	Interval data	N/A
	CareerFat	The student's father's job	Open answer	Discrete data	N/A
	EduMot	The student's mother's educational level	Under high school	UnderHi	2,854 (57.47%)
			High school	Hi	1,245 (25.07%)
			Undergraduate	Uni	417 (8.40%)
			Graduate school	PostGrad	93 (1.87%)
	AgeMot	The student's mother's age	Open answer	Interval data	N/A
	CareerMot	The student's mother's job	Open answer	Discrete data	N/A
	EcoStt	The student's family economic status	Low-income family	poor	379 (7.63%)
			Medium-income family	med	3912 (78.78%)
			Wealthy family	rich	525 (10.57%)
	KnowledgeInc	Whether the student is aware of the family's monthly income	Yes	yes	697 (14.04%)
			No	no	3953 (79.60%)
	EstIncome	Estimated monthly income of the student's family	Open answer	Interval data	N/A

Table A2. Description of variables related to book-reading preferences, book-reading frequency/habits, and classroom activities.

Category	Variable	Variable description	Variable level(s)	Code/Data type	Number (%)
3. Book reading preferences	*Readbook*	Whether the student is interested in reading book or not	Yes	yes	3,515 (70.78%)
			No	no	326 (6.56%)
	Topic	The student's most favorite type of book	Mathematics/physics	a	705 (14.20%)
			Literature	b	1,369 (27.57%)
			Foreign language	c	367 (7.39%)
			Natural science/chemistry/biology	d	934 (18.81%)
			History/geography	e	558 (11.24%)
			Information technology	f	711 (14.32%)
	Typebook	The student's preferred type of book if he/she is gifted, besides textbooks	Novel	a	1,910 (38.46%)
			Biography	b	351 (7.07%)
			Popular science	c	576 (11.60%)
			Arts	d	657 (13.23%)
			Vocational instruction	e	728 (14.66%)
			Other	f	676 (13.61%)
	Reason	The student's reason for choosing that type of book (subsequent variable of "Typebook")	Personal preferences	a	4088 (82.32%)
			Recommended by parents	b	183 (3.69%)
			Recommended by teachers/friends	c	345 (6.95%)
			Serendipity	d	290 (5.84%)
	PrioAct	The student's primary activity when meeting a good book	Sharing with friends/family	a	2,165 (43.60%)
			Recording	b	1,034 (20.82%)
			Applying the content to daily life	c	339 (6.83%)
			Reflecting and relating to personal knowledge	d	1,360 (27.39%)
	AftAct	The student's primary activity after reading a good book	Finding more books on the exact issue	a	2,420 (48.73%)
			Finding more books on the related issue	b	826 (16.63%)
			Finding books on the new issue	c	678 (13.65%)
			Reading the book again	d	955 (19.23%)
	Read_like	The student's two most favorite books	Open answer	Discrete data	N/A

Category	Variable	Variable description	Variable level(s)	Code/Data type	Number (%)
4. Book reading frequency/ habits	*TimeSci*	The amount of time a student spent reading science book daily	Less than 30 minutes	1	2,340 (47.12%)
			Between 30 and 60 minutes	2	2296 (46.23%)
			Over an hour	3	990 (19.94%)
	TimeSoc	The amount of time a student spent reading literature/social science book daily	Less than 30 minutes	1	2898 (58.36%)
			Between 30 and 60 minutes	2	1641 (33.04%)
			Over an hour	3	255 (5.13%)
	Readstory	Whether the student's parents read a story for him/her or not	Yes	yes	1190 (23.96%)
			No	no	3484 (70.16%)
	Buybook	Whether the student's parents buy books for him/her or not	Yes	yes	4134 (83.25%)
			No	no	710 (14.30%)
	Source	The student's main source of book	Buying books on their own or with parents' money	buy	1659 (33.41%)
			Borrowing from friends or libraries	borrow	3140 (63.23%)
			Being gifted or rewarded	gift	120 (2.42%)
5. Class-room activities	*EncourAct*	Whether the student is interested in activities that encourage reading	Yes	yes	4249 (85.56%)
			No	no	557 (11.22%)
	MostlikedAct	The student's most favorite activity that encourages reading	Book exhibition	a	1796 (36.17%)
			Storytelling competition	b	1210 (24.37%)
			Story-writing competition	c	526 (10.59%)
			Illustrating books' content by drawing	d	1092 (21.99%)
	Bookcase	The student's perception of the classroom's bookshelf condition	Diverse and interesting	a	2230 (44.91%)
			Lack of good titles	b	1237 (24.91)
			Lack of book	c	348 (7.01%)
			No bookshelf	d	1006 (20.26%)
	Notread_like	The student's three favorite books that are not available in the classroom's bookshelf	Open answer	Discrete data	N/A

AUTHOR BIOGRAPHY

Quan-Hoang Vuong (Ph.D., Université Libre de Bruxelles) is the director of Centre for Interdisciplinary Social Research, Phenikaa University in Hanoi, Vietnam. He is chairman of the Vietnam Chapter of the European Association of Science Editors and serves in the NAFOSTED Scientific Council on Basic Research in the Social Sciences and Humanities (2019–2021). Dr. Vuong has published more than 160 academic articles and book chapters in about 60 refereed journals and books by such publishers as Elsevier, Inderscience, Nature Publishing Group, Springer, Wiley, and World Scientific.
ORCID: 0000-0003-0790-1576

Viet-Phuong La is a researcher at the Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam, and a software engineer for A.I. for Social Data Lab, Vuong & Associates, Hanoi, Vietnam.
ORCID: 0000-0002-4301-9292

Manh-Toan Ho holds an MA at the National Economics University in Hanoi, Vietnam. He is working as a researcher in the Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam, and he is also a science communicator for EASE Vietnam SciComm/SSHPA (https://sc.sshpa.com/), in addition to other media outlets in Vietnam.
ORCID: 0000-0002-8292-0120

Thanh-Hang Pham holds a Master of Business Administration from School of Business, La Trobe University, Melbourne, Australia, and is currently taking the PhD program in RMIT Vietnam University, Hanoi, Vietnam.
ORCID: 0000-0002-7232-7948

Thu-Trang Vuong is a graduate student at École doctorale of Sciences Po Paris. She has worked on a varied range of topics, including entrepreneurship in emerging countries, public health, education, political analyses of culture and religion, with the aim of providing insights for policy-making.
ORCID: 0000-0002-7262-9671

Ha-My Vuong is a research assistant for A.I. for Social Data Lab, Vuong & Associates, Hanoi, Vietnam.
ORCID: 0000-0002-3058-8015

Minh-Hoang Nguyen holds an MSc in Sustainability Science from Ritsumeikan Asia Pacific University, Beppu, Japan, where he now continues with his PhD track. He works as a researcher in the Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, Vietnam. His research interest is about psychological issues. He believes understanding human perceptions is a fundamental approach for achieving sustainability in multiple disciplines. ORCID: 0000-0002-7520-3844